INSTITUTE OF ACTUARIES OF INDIA

EXAMINATIONS

25th May 2024

Subject CS1A – Actuarial Statistics (Paper A)

Time allowed: 3 Hours 15 Minutes (09.30 – 12.45 Hours)

Total Marks: 100

1. ***Mark allocations are shown in brackets.***

2. ***While attempting questions 1 to 20, you are only required to state the right option number in the designated section in the answer sheet. You are NOT required to give reasons / show supporting calculations to justify your choice.***

Q.1 to Q.3 are based on the information presented below:

A research firm is analysing the number of likes for new posts by the best social media influencers in the country on different social media platforms *viz.* Xstagram, Scapebook and Fritter.

It is estimated that the number of likes received per second for new posts on Xstagram follows a Poisson process with a mean rate of 0.4 likes per second.

Q. 1) A renowned social media influencer Mr. Numero Uno has openly claimed that his new reel on Xstagram will receive exactly 40 likes in a minute. Probability that this claim turns out to be true is –

A. 0.193%
B. 6.295%
C. 8.115%
D. 0.075% **[2]**

Q. 2) The mean waiting time between two consecutive likes for a new post on Xstagram will be –

A. 4 seconds
B. 2.5 seconds
C. 5 seconds
D. 0.4 seconds **[2]**

Q. 3) A novice on social media, Mr. Numero Cero has already waited for 10 seconds for receiving a like for his new post on Xstagram with no success. What is the probability that he will have to wait for at least 10 more seconds for receiving his first like?

Hint: *Use the memoryless property of an exponential distribution.*

A. 0.005%
B. 0.034%
C. 1.832%
D. 2.732% **[2]**

Q.4 and Q.5 are based on the information presented below:

Two other social media platforms viz. Scapebook and Fritter are being compared by the research firm in terms of the waiting time between two consecutive likes. Target audience on both these platforms is different – Scapebook is used mostly by youngsters up to age 40 and Fritter is generally used by the older population with ages 40 and above.

Waiting times are modelled using independent exponential variables X (Scapebook) and Y (Fritter) with means of 1 second and 2 seconds respectively.

Q.4) Which of the following is the correct expression representing f (x, y) i.e. the joint distribution of X and Y?

A. $f(x, y) = \exp(-x) \times \frac{1}{2} \exp(-\frac{1}{2}y)$ for $\infty > x, y > 0$;
$= 0$ otherwise
B. $f(x, y) = 2 \exp(-2x) \times \exp(-y)$ for $\infty > x, y > 0$;
$= 0$ otherwise

C. $f(x, y) = \exp(-x) \times 2\exp(-2y)$ for $\infty > x, y > 0$;
$= 0$ otherwise
D. $f(x, y) = \frac{1}{2}\exp(-\frac{1}{2}x) \times \exp(-y)$ for $\infty > x, y > 0$;
$= 0$ otherwise **[2]**

Q. 5) Let the joint MGF $M_{x, y}(t, s)$ of the joint distribution of X and Y be defined as follows:

$$M_{x, y}(t, s) = E[\exp(xt + ys)].$$

Which of the following statements is TRUE in respect of the joint MGF of random variables X and Y i.e. $M_{x, y}(t, s)$?

A. $M_{x, y}(t, s) = M_x(t) + M_y(s)$
B. $M_{x, y}(t, s) = M_x(t) \times M_y(s)$
C. $M_{x, y}(t, s) = M_x(t) - M_y(s)$
D. None of the above **[2]**

Q.6 to Q.10 are based on the information presented below:

The number of cheques dishonoured y, in a batch of x cheques meant for clearing in a local branch of a cooperative bank is modelled as a Poisson random variable with mean λx, where λ is unknown.

Data is available from six independent batches of proposals as follows:

Batch No.	1	2	3	4	5	6
x:	50	75	105	150	40	60
y:	1	21	17	9	4	3

$\sum x = 480$; $\sum x^2 = 46850$; $\sum y = 55$; $\sum y^2 = 837$; $\sum xy = 5100$

Q. 6) The method of moments estimate $\hat{\lambda}_{MOM}$ based on the given data is equal to –

A. 8.727
B. 0.786
C. 0.115
D. 1.271 **[1]**

Q.7) Let us define the least square estimator $\hat{\lambda}_{LSE}$. By definition, it is that value of λ for which $\sum(y - E(y))^2$ is minimized.

Which of the following represents the correct expression for the least square estimator of λ i.e. $\hat{\lambda}_{LSE}$?

A. $\hat{\lambda}_{LSE} = \frac{\sum y^2}{\sum x^2}$
B. $\hat{\lambda}_{LSE} = \frac{\sum xy}{\sum y^2}$
C. $\hat{\lambda}_{LSE} = \frac{\sum x}{\sum y}$
D. $\hat{\lambda}_{LSE} = \frac{\sum xy}{\sum x^2}$ **[3]**

Q. 8) Which of the following is the likelihood function of λ?

A. $L(\lambda) = e^{-\lambda \sum x} \times \prod(\lambda x)^y \times c$
B. $L(\lambda) = e^{-\lambda \prod x} \times \prod(\lambda x)^y \times c$
C. $L(\lambda) = e^{-\lambda \sum x} \times \sum(\lambda x)^y \times c$
D. $L(\lambda) = e^{-\lambda \prod x} \times \sum(\lambda x)^y \times c$

[2]

Q. 9) Which of the following is TRUE regarding the maximum likelihood estimator $\hat{\lambda}_{MLE}$ which maximizes the likelihood function as selected in Q.8 –

A. $\hat{\lambda}_{MLE} = \hat{\lambda}_{LSE}$
B. $\hat{\lambda}_{MLE} = \hat{\lambda}_{MOM}$
C. $\hat{\lambda}_{MLE} = \hat{\lambda}_{LSE} = \hat{\lambda}_{MOM}$
D. None of the above

[3]

Q. 10) The least square estimate $\hat{\lambda}_{LSE}$ based on the given sample data is –

A. 0.0179
B. 6.0932
C. 8.7273
D. 0.1089

[1]

Q.11 and Q.12 are based on the information presented below:

Following data tabulates the number of matches (in thousands) on a dating application Rumble for the last five years for four different categories.

		Number of matches in Year (j)						
		2019	**2020**	**2021**	**2022**	**2023**	**Mean**	**Variance**
Category (i)	**Under 18**	48	53	42	50	59	50.4	39.3
	Adults	64	71	64	73	70	68.4	17.3
	Divorced	85	54	76	65	90	74.0	215.5
	Seniors	44	52	69	55	71	58.2	132.7
				Variance of Means:			**110.57**	

The credibility factors Z_i for all categories are to be determined using the assumptions of EBCT (Empirical Bayes Credibility Theory) Model 1.

Q. 11) Value of credibility factor Z_A (for adults category) is –

A. 0.8176
B. 0.8485
C. 0.7812
D. 0.8169

[3]

Q. 12) Value of expected number of matches for adults category (in thousands) is –

A. 68.40
B. 67.37
C. 62.75
D. 63.81

[2]

Q.13 to Q.15 are based on the information presented below:

A 4 × 4 correlation matrix using Karl Pearson's Method has been constructed for the data collected from the dating application Rumble tabulating the sample correlation coefficients:

	Under 18	**Adults**	**Divorced**	**Seniors**
Under 18	1.00	0.63	0.12	0.13
Adults	0.63	1.00	-0.51	0.02
Divorced	0.12	-0.51	1.00	0.33
Seniors	0.13	0.02	0.33	1.00

Q. 13) Based on this table, which of the following statements is FALSE?

A. Teenagers(under 18) and adults tend to show similar dating behaviour.
B. Dating behaviour for seniors is hardly related to the dating behaviour for adults.
C. Adults and divorcees tend to show dissimilar dating behaviour.
D. Dating behaviour of seniors is more similar to teenagers (under 18) than to divorcees. **[1]**

Q. 14) It is decided to perform a statistical test to check whether the dating behaviour of teenagers (under 18) and divorcees is uncorrelated with each other. Students' t distribution is to be used for performing this statistical test.

What is the value of the test statistic for the above test?

A. 0.27
B. 3.70
C. 0.21
D. 4.78 **[2]**

Q. 15) Which one of the following options correctly represents the p-value for this test and the conclusion of this statistical test at 5% level of significance?

	p-value	**Conclusion of the statistical test**
A.	< 40%	Dating behaviour of teenagers (under 18) and divorcees is uncorrelated
B.	< 5%	Dating behaviour of teenagers (under 18) and divorcees is correlated
C.	> 40%	Dating behaviour of teenagers (under 18) and divorcees is uncorrelated
D.	< 1%	Dating behaviour of teenagers (under 18) and divorcees is correlated

[2]

Q.16 to Q.20 are based on the information presented below:

Your team member Mr. Left was working on a project which involved fitting a linear regression model for testing effectiveness of new drug of a fertilizer manufacturing company. The company is launching a new fertilizer DAP and is testing the impact of the fertilizer on the crop yield based on trials done over 20 hectares of paddy crop. The model helps to predict the crop yield(Y) of paddy crop per hectare based on amount of fertilizer(X) employed in the field per hectare.

Mr. Left has resigned and the fitted linear regression model which was saved on his laptop cannot be retrieved as his laptop has been formatted. You have to give a presentation to the fertilizer manufacturing company tomorrow. On his writing desk you find a printout which contains details of a sample of ten X_i values, corresponding fitted $\hat{Y}_i$ values.

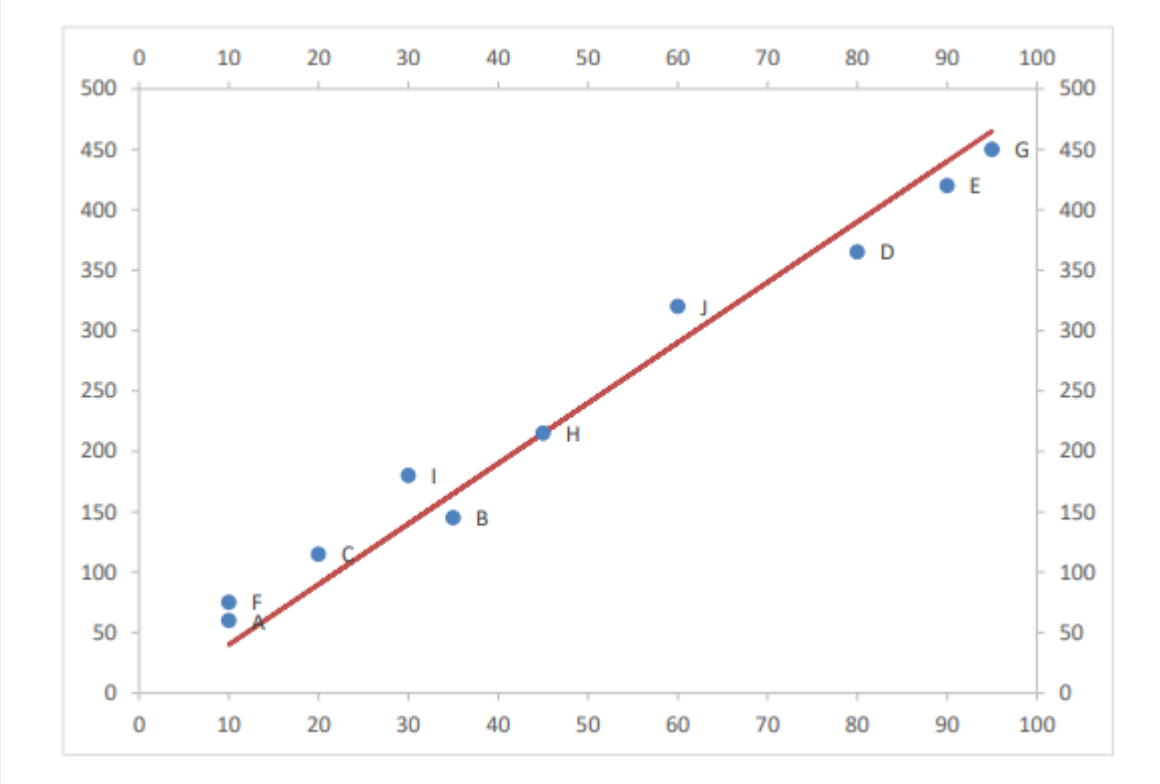
The ten pairs of of $(X_i, \hat{Y}_i)$ from the printout are as follows:

A = (10, 40); B = (35, 165); C = (20, 90); D = (80, 390); E = (90, 440);
F = (10, 40); G = (95, 465); H = (45, 215); I = (30, 140); J = (60, 290).

You had called Mr. Left and he informed you that the model was fitted based on 100 values of X and Y obtained from the trails conducted by the company. Mr. Left also informed you that he remembers that the coefficient of determination i.e. R^2 for this model is 70%.

Apart from the values of $(X_i, \hat{Y}_i)$, the printout also contained information about residual values e_i for each pair A to J. Based on this, you have constructed the following scatter plot. Every point shown in the plot is (X, Y) where $Y = \hat{Y} + e$.

Also, the fitted regression line has been back-calculated based on the data available from the printout and plotted in the graph as the red line.



Q. 16) How many minimum pairs of $(X_i, \hat{Y}_i)$ will you require to obtain the regression equation of Y on X: $Y = \alpha + \beta \times X$?

A. 2
B. 3
C. 10
D. 1

[2]

Q. 17) Based on the pairs A to F found in the printout, the regression equation for Y on X is –

A. Y = – 5 + 10X
B. Y = 5 + 10X
C. Y = 10 + 5X
D. Y = – 10 + 5X. **[2]**

Q. 18) Using the plot depicted above, which one of the following options represents possible values for $\overline{X}$ and $\overline{Y}$ for the trial data set of 100 values?

A. $\overline{X}$ = 47.50 and $\overline{Y}$ = 227.50
B. $\overline{X}$ = 35.00 and $\overline{Y}$ = 165.00
C. $\overline{X}$ = 30.00 and $\overline{Y}$ = 140.00
D. $\overline{X}$ = 45.00 and $\overline{Y}$ = 215.00

Hint: *Regression line of Y on X passes through the point* $(\bar{X}, \bar{Y})$. **[2]**

Q. 19) What will be the value of Adjusted R^2 for this bivariate regression model?

A. 66.25%
B. 70.00%
C. 69.69%
D. 70.14% **[2]**

Q. 20) You decide to check whether the parameters are significant (i.e. non-zero) using one way ANOVA (analysis of variance).

If the value of ANOVA F-statistic is calculated to be 229, what is your conclusion at 1% level of significance?

A. $\beta = 0$
B. $\beta \neq 0$
C. $\alpha \neq 0$
D. $\alpha = 0$ **[2]**

Q. 21) A pair of two biased two-sided coins is tossed once. The random variables representing the outcome on the first and the second coin are denoted by X and Y respectively. X and Y take numerical values of 1 and 0 if the outcome is Heads (H) and Tails (T) respectively.

These two random variables X and Y, have the following joint probability function:

		X (Outcome on first coin)	
		Heads (H)	**Tails (T)**
Y (Outcome on second coin)	**Heads (H)**	P (X = "H", Y= "H") = 2/9	P (X = "T", Y= "H") = 4/9
	Tails (T)	P (X = "H", Y= "T") = 1/9	P (X = "T", Y= "T") = 2/9

i) Determine conditional expectation: E(Y | X = "H"). (1)

ii) Determine conditional variance: Var(Y | X = "H"). (2)

iii) State the probability functions of the marginal distributions of X and Y. (2)

iv) Determine E(Y) and Var(Y) using the marginal distribution of Y as determined in part (iii). (1)

v) Based on the marginal distributions of X and Y as determined in part (iii), determine which coin is biased in favour of Heads (H) and which coin is biased in favour of Tails (T)? (1)

vi) Check whether the random variables X and Y are independent of each other and suitably comment on your findings in parts (i), (ii) and (iv). (3)

[10]

Q. 22) Total processing time for a unit of a catalytic converter for a car is normally distributed with a mean of 20 minutes and a standard deviation of 5 minutes.

Let us consider a random sample of 5 units of the catalytic converter which comprises of X_1, X_2, X_3, X_4 and X_5 which are independent and identically distributed random variables from the normal distribution as mentioned above.

i) Determine the probability that the sample mean of the processing time ($\bar{X}$) for these 5 units is less than 15 minutes. (2)

ii) Determine the probability that the sample standard deviation of the processing time (S) for these 5 units is greater than 6.65 minutes. (2)

iii) Determine the probability that both sample mean of processing time ($\bar{X}$) is less than 15 minutes and sample standard deviation of processing time (S) is greater than 6.65 minutes for this sample of 5 units of the catalytic converter. (2)

Random variable $[(n - 1) * S^2 / \sigma^2]$ is distributed as a chi-square variable with (n – 1) degrees of freedom.

Mean of this random variable is (n – 1) and its variance is 2 * (n – 1).

iv) In light of the above information, calculate the value of $E(S^2)$ and $Var(S^2)$ for the random sample of 5 units of the catalytic convertor. If the sample size is increased to 100 units instead of 5 units, what will be the impact on the value of $E(S^2)$ and $Var(S^2)$? Briefly comment on the same. (4)

[10]

Q. 23) You are working in a firm of climate risk actuaries and your firm has developed an Actuarial Climate Risk Index (ACRI) for your country which is determined based on five key elements viz. temperature, high precipitation, drought, strong winds and sea level.

A high value of the index is associated with greater climate risk for the country.

Over the past few years, an increase in the value of the index has been consistently observed. Every year the value of this index is increasing by one as compared to the previous year. You have been given the responsibility to check whether there is any association between ACRI and climate related deaths in the country.

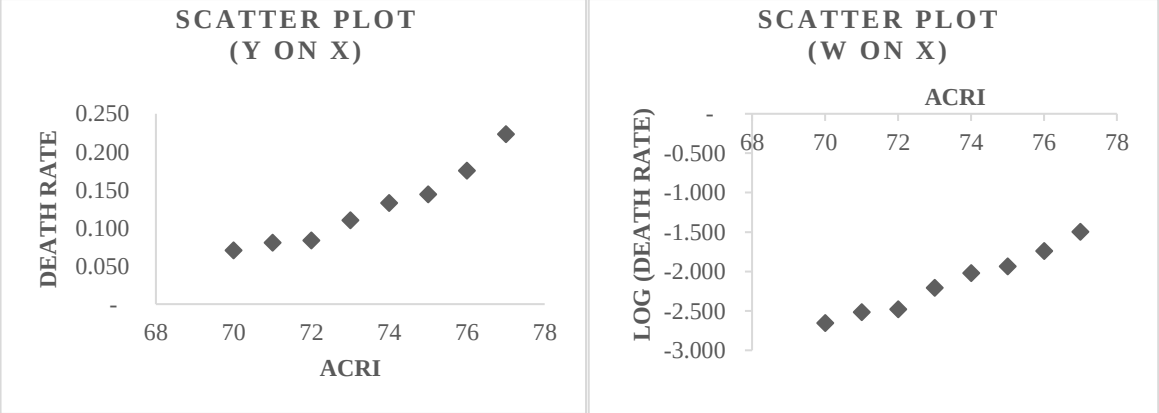
The table below gives year-wise details of ACRI represented by x and the number of deaths due to climate change represented by n. The population exposed to climate change related events, denoted E, have also been given. The values of death rates y, where y = n / E, and the log (death rates), denoted w, are also given.

Year	ACRI (x)	Number of deaths (n)	Exposure (E)	y = n / E	w = log(y)
2016	70	30	426	0.07042	− 2.6532
2017	71	38	471	0.08068	− 2.5173
2018	72	38	454	0.08370	− 2.4805
2019	73	53	482	0.10996	− 2.2077
2020	74	59	445	0.13258	− 2.0205
2021	75	61	423	0.14421	− 1.9365
2022	76	82	468	0.17521	− 1.7417
2023	77	96	430	0.22326	− 1.4994

$\sum x = 588$; $\sum x^2 = 43260$; $\sum w = -17.0568$; $\sum w^2 = 37.5173$; $\sum xw = -1246.7879$

A regression line is to be fit with death rates as the response variable and ACRI as the explanatory variable. However, your teammates have suggested that logarithm of the death rates i.e. w should be used instead of y.

Following scatter plots have been obtained for the data.



i) Based on the scatter plots as depicted above, briefly explain why a logarithmic transformation could be necessary in this case. (2)

ii) Calculate the least squares fit regression line in which w is modelled as response variable and x is modelled as explanatory variable. (6)

iii) Calculate the fitted value for the number of deaths in the year 2018. (2)

[10]

Q. 24) Let X be the performance rating that an employee gets in Math Solutions Ltd. X is modelled using normal distribution i.e. $X \sim N(\mu, \sigma^2)$ where μ is unknown. The parameter μ follows a uniform distribution over the interval (0, 10). Data for 150 employees who were part of the current year's performance appraisal exercise is collected and the average rating for these 150 employees is found out to be 5.

We want to find the posterior density of μ for which following workings have been done.

Step1: Prior density of μ

$f(\mu) = 1 / (A - B) = 1/10 = \text{constant}.$

Step 2: Likelihood function of X ignoring any coefficient of proportionality:

$L(x, \mu) = \text{constant} * e^{\left\{\frac{-1}{2C} \times \sum(x - D)^2\right\}}.$

Step 3: Arriving at the posterior density of μ:

$f(\mu \mid x)$

$= f(\mu) * L(x, \mu)$

$= \text{constant} * e^{\left\{\frac{-1}{2C} \times \sum(x - D)^2\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{2C} \times (\sum x^2 - 2\mu \sum x + n\mu^2)\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{2C} \times (-2\mu \sum x + E)\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{\frac{2\sigma^2}{F}} \times (-2\mu\bar{x} + \mu^2)\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{\frac{2\sigma^2}{F}} \times (\mu^2 - 2\mu\bar{x} + G)\right\}}$

$= \text{constant} * e^{\left\{\frac{-1}{2H} \times (\mu - I)^2\right\}}$

Hence, the posterior density of μ is as follows:

$\mu \sim N(5, \sigma^2 / J).$

i) Determine the values of the missing terms A, B, C, D, E, F, G, H, I and J. You are NOT required to provide any justification / supporting calculations. (5)

ii) Determine the Bayesian estimate for μ under –

a) Squared error loss;
b) All-or-nothing loss;
c) Absolute error loss. (2)

iii) Assuming σ^2 to be 25, determine a 95% equal-tailed credible interval for μ.

Hint: *An equal-tailed credible interval is a confidence interval determined using the posterior distribution of μ.* (2)

An alternative to equal-tailed credible interval is highest posterior density interval for μ. This interval is such that the minimum density of any point within this interval is equal to or higher than the density outside this interval.

iv) Without performing any additional calculations, state the 95% highest posterior density interval for μ. Justify your answer. (1)

[10]

Q. 25) Recently the results of December 2023 Diet Examinations conducted by the Institute of Actuaries of Statistica (IAS) were declared. Following news was published in one of the local newspaper:

"Candidates from Region X outperform candidates from Region Y in the December 2023 actuarial examinations with pass rates of 52% and 40% respectively."

Your manager, Mrs. Numara is a reputed actuary in the industry and she belongs to Region Y. She was not convinced with such a significant difference in pass rates for Region X and Region Y. She has asked you to contact the officials of IAS and statistically test whether candidates from Region X are smarter than candidates from Region Y when it comes to passing actuarial examinations.

You reached out to officials of IAS and learned that 2050 candidates from Region X and 800 candidates from Region Y appeared in the examinations, with reported pass rates of 52% and 40% respectively, as reported in the newspaper. You have opted to use a non-parametric method for conducting the statistical analysis.

i) Name any two non-parametric approaches to hypothesis testing. (1)

You decided to use contingency tables for performing this statistical test.

ii) State the null hypothesis (H_0) and alternate hypothesis (H_1) for this test. (1)

iii) Complete the observed frequencies table based on the data collected from the officials of IAS. Write the values of XP_O, XF_O, YP_O and YF_O in the answer script. You are NOT required to copy the table and NOT required to show any supporting calculations.

Observed Frequencies

	Region X	**Region Y**	**Total**
Pass	XP_O	YP_O	
Fail	XF_O	YF_O	
Total			

(1)

In the expected frequencies table corresponding to the above observed frequencies, let us define four quantities *viz*. XP_E, XF_E, YP_E and YF_E representing candidates from Region X expected to pass, candidates from Region X expected to fail, candidates from Region Y expected to pass and candidates from Region Y expected to fail respectively.

iv) Write the values of XP_E, XF_E, YP_E and YF_E in the answer script (rounded off to the nearest integer). You are NOT required to show any supporting calculations. (2)

v) Calculate the value of the χ^2 test statistic. Based on the value of the test statistic, show that the above test in fact strengthens the claim that candidates from Region X are smarter than candidates from Region Y when it comes to passing actuarial examinations. Test at 0.5% level of significance. (3)

Your manager is still not convinced with the results of the statistical test and makes a request to the Examinations Controller of IAS under the Right to Information (RTI) Act asking for detailed subject-wise and region-wise results data for December 2023 Diet Examinations.

Following data is obtained through RTI.

	Region X			Region Y		
Subject	**Appeared**	**Passed**	**Pass %**	**Appeared**	**Passed**	**Pass %**
CS1	500	250	50%	100	60	60%
CM1	600	300	50%	50	40	80%
CB1	400	250	63%	100	90	90%
CB2	450	250	56%	50	35	70%
CM2	50	10	20%	200	50	25%
CS2	50	6	12%	300	45	15%
Total	**2050**	**1066**	**52%**	**800**	**320**	**40%**

Based on the above data, your manager has criticized the conclusion you presented in part (v) of your analysis, claiming that your statistical test led to an incorrect conclusion. She argues that across all subjects, the pass rate for Region Y consistently exceeds that of Region X based on the data provided.

vi) Defend your test results obtained in part (v) clearly addressing the observation made by Mrs. Numara. You are NOT required to perform any additional calculations. (2)

[10]

Q. 26) Railway Ticketing Corporation of Actuaria (RTCA) is an online ticketing platform for booking tickets for railway travels all throughout the country. It has recently started displaying probability of confirmation (p_c) in case of waiting list tickets.

Past data is available for the following explanatory variables:

Wknd : a categorical variable with value = 1 in case if it is a weekend and value = 0 in case if it is a weekday

Kms : a numerical variable which captures the distance between the boarding station and the destination in kms

This probability p_c is being determined by using a binomial generalized linear model (GLM) with the canonical link function. The linear predictor has the form:

$$g(p_c) = \alpha + \beta_{w\ i=0,1} + \beta_k * \text{kms}$$

where: $\beta_{w\ i=0}$ is used in case of a weekday and $\beta_{w\ i=1}$ is used in case of a weekend

The analysis of the past data gave the following estimates for the model:

Coefficient	Estimate	Standard Error
α	- 0.372	0.053
$\beta_{w\ i=0}$	0.036	0.023
$\beta_{w\ i=1}$	- 0.100	0.080
β_k	- 0.003	0.001

i) Determine whether the coefficients $\beta_{w\ i=0}$, $\beta_{w\ i=1}$ and β_k are significant by comparing the estimate of these coefficients with their respective standard errors. You are NOT required to calculate p-values. (2)

ii) Show that the probability p_c that a waiting list ticket will be confirmed for a distance of 300 kms when the travel is on a weekend is 0.20. Use the canonical link function relevant to a binomial model. (3)

Mr. Traveller travels between two cities *viz*. "Thousand Palms" and "Forty Miles" every weekend. Distance between these two cities is 300 kms. He books the tickets invariably late in the evenings on Friday and always features in the waiting list. In case if his ticket is not confirmed then he needs to travel by a private bus.

iii) Calculate the probability that Mr. Traveller travels at least 2 times by train during the coming month. Use the estimate of p_c from part (ii). Assume that a month has 4 weeks. (3)

iv) State how the linear predictor as given above will change if an interaction term between the covariates (wknd * kms) is also included in the model. (2)

[10]

****